Available online at www.sciencedirect.com

ScienceDirect

Physics of Life Reviews 11 (2014) 598–618

PHYSICS of LIFE reviews

www.elsevier.com/locate/plrev

Review

Approaching human language with complex networks

Jin Cong [a], Haitao Liu [a,b,*]

[a] *School of International Studies, Zhejiang University, Hangzhou, CN-310058, China*
[b] *Ningbo Institute of Technology, Zhejiang University, Ningbo, CN-315100, China*

Received 16 March 2014; received in revised form 7 April 2014; accepted 15 April 2014

Available online 18 April 2014

Communicated by L. Perlovsky

Abstract

The interest in modeling and analyzing human language with complex networks is on the rise in recent years and a considerable body of research in this area has already been accumulated. We survey three major lines of linguistic research from the complex network approach: 1) characterization of human language as a multi-level system with complex network analysis; 2) linguistic typological research with the application of linguistic networks and their quantitative measures; and 3) relationships between the system-level complexity of human language (determined by the topology of linguistic networks) and microscopic linguistic (e.g., syntactic) features (as the traditional concern of linguistics). We show that the models and quantitative tools of complex networks, when exploited properly, can constitute an operational methodology for linguistic inquiry, which contributes to the understanding of human language and the development of linguistics. We conclude our review with suggestions for future linguistic research from the complex network approach: 1) relationships between the system-level complexity of human language and microscopic linguistic features; 2) expansion of research scope from the global properties to other levels of granularity of linguistic networks; and 3) combination of linguistic network analysis with other quantitative studies of language (such as quantitative linguistics).

Keywords: Human language; Complex networks; Network topology; Linguistics; Linguistic typology

1. Network models and measures of human language

We live in a world pervaded by networks, i.e., systems which can be represented by graphs, with the system elements as vertices (nodes) and the relations between the elements as edges (links) [1,2]. The great majority of real-world networks (biological, social, technological, etc.) are *complex networks* [3], which are neither regular (as in the case of regular lattices) nor random (with any pair of vertices having a fixed probability to be linked) [4, 5] and exhibit emergent properties which cannot be inferred on the basis of their component parts [6, p. 47]. The recent decade has witnessed the boom of networks science and an explosion of interest in complex networks across

* Corresponding author at: Ningbo Institute of Technology, Zhejiang University, No. 1 Xuefu Road, Ningbo, Zhejiang, CN-315100, PR China. Tel.: +86 571 88981923.
E-mail address: lhtzju@gmail.com (H. Liu).

http://dx.doi.org/10.1016/j.plrev.2014.04.004
1571-0645/

a multitude of disciplines ranging from natural sciences to social sciences and humanities [1,2,7–17]. In complement to the reductionist approach as commonly used in modern science, this new science of networks makes it possible to probe into the complexity of real-world systems in their entirety and thus constitutes one, if not the only, solution to the challenge of "reassembling" complex systems and capturing their holistic properties [18, p. 93]. Indebted substantially to graph theory and statistical physics, the models and quantitative tools employed by networks science provide a unifying framework for the structure and dynamics of real-world networks of various natures and thus facilitate communication between different disciplines.

Language is "one of the wonders of the natural world" [19, p. 15] and "what makes us human" [20, p. 4]. Recognition is increasing that human language can also be modeled and analyzed with complex networks [17,21–29]. Among the growing enthusiasm for complex networks in recent years, the inquiry into human language from the complex network approach has arisen as a highly productive area, which is characterized by the convergence of disciplines such as statistical physics, systems science, linguistics, cognitive science, and natural language processing. This interdisciplinary endeavor contributes both methodological and substantive insights into human language as a system. That language is a system is a central assumption of modern linguistics [30, p. 1]. According to Saussure, the father of modern linguistics, language is a system in which each linguistic unit is defined by, and only by, its relations with the other units [31]. The Saussurean conception of language as a system is generally consistent with the modern definition of a system [32] and is manifested to varying degrees in a number of subsequent linguistic theories and schools (e.g. [33–38]).

In the absence of operational methodology, the system perspective on human language has not been carried any further and only amounts to a metaphor. Instead, linguists are preoccupied with detailed structural features of human language, which can be easily handled with the reductionist approach. Inspired by complexity theory, it is recently acknowledged that language is a complex system [39,40]. If language is conceived of as a complex system of linguistic units and their relations, it is expected to exhibit emergent properties at the system-level due to the microscopic-level interactions between the system elements. Complex networks provide appropriate modeling for human language as a complex system and powerful quantitative measures for its complexity at the system-level. The flourishing research of linguistic networks has introduced a holistic and quantitative approach to the understanding of human language as a system. In addition, the unifying framework of complex networks places linguistic research in a broader and interdisciplinary context. This context is what linguistic research intrinsically deserves.

Appropriate use of network analysis depends on the right choice of network representation [41]. Linguistic networks are network models for human language as a system. As there is no one network model which can cover the multi-faceted nature of human language, researchers rely on network models of various language sub-systems, each of which is a particular aspect or level of language. Like the network model of any other type of system, the basic form of a linguistic network N is a pair of sets $N = (V, E)$, whereby V is the set of vertices representing the linguistic units and E the set of edges representing the pairwise relations of a particular type between these linguistic units in the language sub-system in question.

Some language sub-systems are inventories of linguistic units (such as words, morphemes and phonemes) as found in a dictionary. For instance, the inventory of words (usually termed as *lexicon*) of a language may be organized by semantic relations (hyponymy, meronymy, antonymy, synonymy, etc.) between the words. Language sub-systems as inventories of linguistic units are modeled by *static linguistic networks* [42]. Analysis of static linguistic networks can shed light on the complex organization of different inventories of linguistic units of human language. A representative example of static linguistic networks is *static semantic networks* [43–48], which are based on such resources as word associations, WordNet, thesaurus, and Semantic Web. A static semantic network models the lexicon of a language, with the words as vertices and their semantic relations as edges. Other static linguistic networks may model language sub-systems pertaining to the formation of particular linguistic units. A network of this type can be constructed so that two linguistic units (such as morphemes and phonemes) as vertices are joined by an edge if they form a larger linguistic unit (e.g., a word) [49–54]. Another way is to capture the similarities of the linguistic units in terms of formation so that two linguistic units (such as words) as vertices are joined by an edge if they are (morphologically or phonologically) similar [55–57].

Other language sub-systems, on the other hand, are those of linguistic units and their relations as found in actual language use. These language sub-systems are modeled by *dynamic linguistic networks* [42]. Dynamic linguistic networks, unlike static linguistic networks, are based on naturally-occurring language data and thus can reflect the complexity of actual language use. The relations between linguistic units in actual language use can be observed at